RESEARCHER DEVELOPMENT

Clustering

(unsupervised learning)

JJ Valletta

February 24, 2015

www.exeter.ac.uk/as/rdp/

Overview

- What is clustering?
- Major types of clustering methods
- k-means clustering
- Agglomerative hierarchical clustering
- Gaussian mixture models
- How do we determine the correct number of clusters?

Overview

- What is clustering?
- Major types of clustering methods
- k-means clustering
- Agglomerative hierarchical clustering
- Gaussian mixture models
- How do we determine the correct number of clusters?

Overview

- What is clustering?
- Major types of clustering methods
- k-means clustering
- Agglomerative hierarchical clustering
- Gaussian mixture models
- How do we determine the correct number of clusters?

Overview

- What is clustering?
- Major types of clustering methods
- k-means clustering
- Agglomerative hierarchical clustering
- Gaussian mixture models
- How do we determine the correct number of clusters?

Overview

- What is clustering?
- Major types of clustering methods
- k-means clustering
- Agglomerative hierarchical clustering
- Gaussian mixture models
- How do we determine the correct number of clusters?

Overview

- What is clustering?
- Major types of clustering methods
- k-means clustering
- Agglomerative hierarchical clustering
- Gaussian mixture models
- How do we determine the correct number of clusters?

Overview

- What is clustering?
- Major types of clustering methods
- k-means clustering
- Agglomerative hierarchical clustering
- Gaussian mixture models
- How do we determine the correct number of clusters?

What is clustering?

What is clustering?

Formal definition

Identifying homogeneous and well separated groups of data points (features) by some similarity measure

What is clustering?

Formal definition

Identifying homogeneous and well separated groups of data points (features) by some similarity measure

Informal definition

The process of stereotyping your data
e.g *these* are round(ish) faces, *these* are short(ish) people

What is clustering?

Formal definition

Identifying homogeneous and well separated groups of data points (features) by some similarity measure

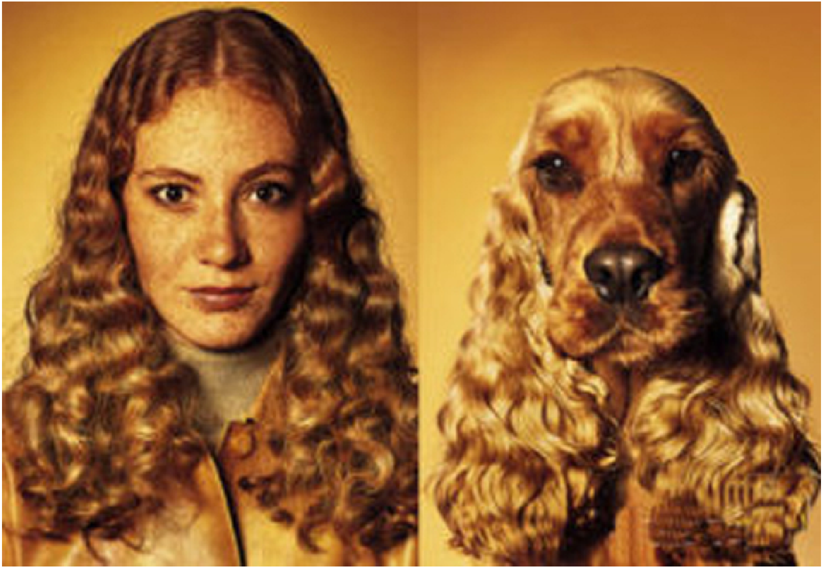
Informal definition

The process of stereotyping your data
e.g *these* are round(ish) faces, *these* are short(ish) people

But how many groups?

An unsolved problem. Issue lies in the subjectivity of the word **similar** and its mathematical definition

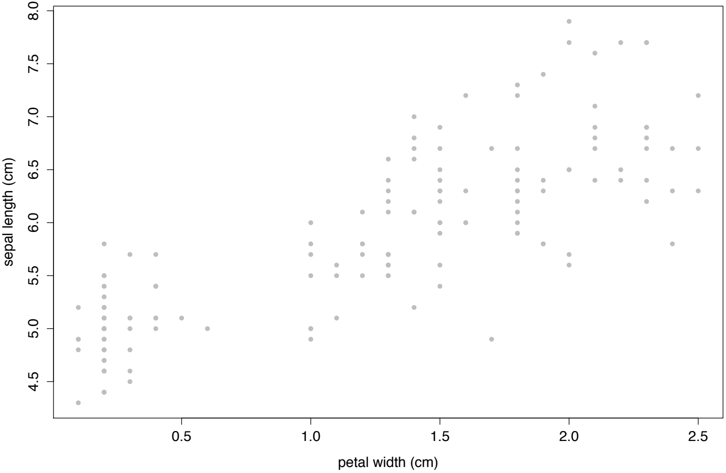
Are they similar?



Are they similar?

What are we after?

- High intra-cluster similarity
- Low inter-cluster similarity
- Elucidate on how the data is structured (maybe identify outliers)



Where is clustering used?

Biological systematics: finding organisms sharing similar attributes

Computer vision: segmenting a digital image for object recognition

Epidemiology: identifying geographical clusters of diseases

Gene expression: discovering co-regulated genes

Medical imaging: differentiating between tissues

Mathematical chemistry: grouping compounds by topological indices

Clustering is particularly useful in applications where labelling the data is very time consuming/expensive

Major types of clustering methods

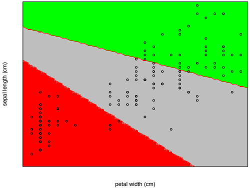
Partitional: The data (feature) space is partitioned into k regions

Hierarchical: Iteratively merging small clusters into larger ones (*agglomerative*) or breaking large clusters into smaller ones (*divisive*)

Distribution-based: Fit k multivariate statistical distributions

Major types of clustering methods

Partitional: The data (feature) space is partitioned into k regions

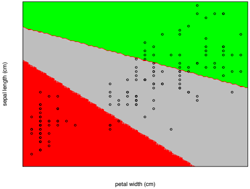


Hierarchical: Iteratively merging small clusters into larger ones (*agglomerative*) or breaking large clusters into smaller ones (*divisive*)

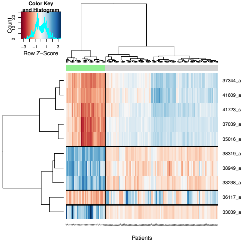
Distribution-based: Fit k multivariate statistical distributions

Major types of clustering methods

Partitional: The data (feature) space is partitioned into k regions



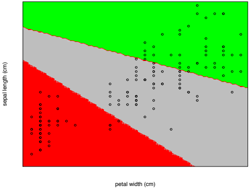
Hierarchical: Iteratively merging small clusters into larger ones (*agglomerative*) or breaking large clusters into smaller ones (*divisive*)



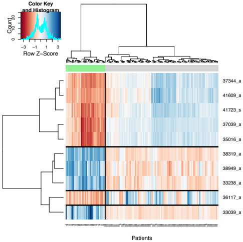
Distribution-based: Fit k multivariate statistical distributions

Major types of clustering methods

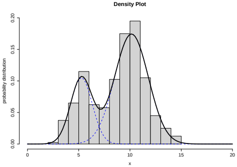
Partitional: The data (feature) space is partitioned into k regions



Hierarchical: Iteratively merging small clusters into larger ones (*agglomerative*) or breaking large clusters into smaller ones (*divisive*)

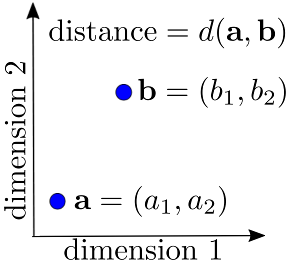


Distribution-based: Fit k multivariate statistical distributions



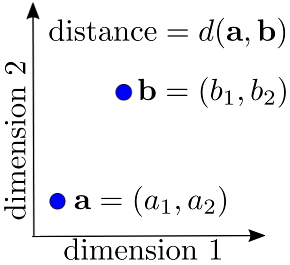
Similarity measures

What is the distance $d(\mathbf{a},\ \mathbf{b})$ between $\mathbf{a}$ and $\mathbf{b}$?


distance $= d(\mathbf{a}, \mathbf{b})$
$\mathbf{b} = (b_1, b_2)$
$\mathbf{a} = (a_1, a_2)$
dimension 2
dimension 1

Similarity measures

What is the distance $d(\mathbf{a},\ \mathbf{b})$ between $\mathbf{a}$ and $\mathbf{b}$?


distance $= d(\mathbf{a}, \mathbf{b})$
$\mathbf{b} = (b_1, b_2)$
$\mathbf{a} = (a_1, a_2)$
dimension 2
dimension 1

example: $\mathbf{a},\ \mathbf{b} \in \mathbb{R}^2$

- **Manhattan**
  $|a_1 - b_1| + |a_2 - b_2|$
- **Euclidean**
  $\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$
- **Minkowski (p-norm)**
  $\sqrt[p]{|a_1 - b_1|^p + |a_2 - b_2|^p}$

in general: $\mathbf{a},\ \mathbf{b} \in \mathbb{R}^d$

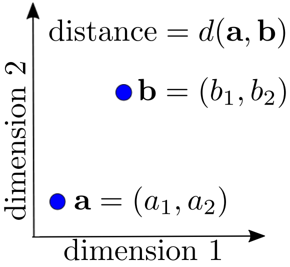
$$\sum_{i=1}^{d} |a_i - b_i| = \|\mathbf{a} - \mathbf{b}\|_1$$

$$\left(\sum_{i=1}^{d} (a_i - b_i)^2\right)^{\frac{1}{2}} = \|\mathbf{a} - \mathbf{b}\|_2$$

$$\left(\sum_{i=1}^{d} |a_i - b_i|^p\right)^{\frac{1}{p}} = \|\mathbf{a} - \mathbf{b}\|_p$$

Similarity measures

What is the distance $d(\mathbf{a},\ \mathbf{b})$ between $\mathbf{a}$ and $\mathbf{b}$?



distance $= d(\mathbf{a}, \mathbf{b})$
$\mathbf{b} = (b_1, b_2)$
$\mathbf{a} = (a_1, a_2)$
dimension 2
dimension 1

example: $\mathbf{a},\ \mathbf{b} \in \mathbb{R}^2$

- **Manhattan**
  $|a_1 - b_1| + |a_2 - b_2|$
- **Euclidean**
  $\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$
- **Minkowski (p-norm)**
  $\sqrt[p]{|a_1 - b_1|^p + |a_2 - b_2|^p}$

in general: $\mathbf{a},\ \mathbf{b} \in \mathbb{R}^d$

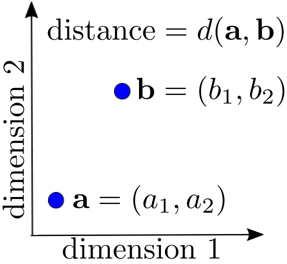
$$\sum_{i=1}^{d} |a_i - b_i| = \|\mathbf{a} - \mathbf{b}\|_1$$

$$\left(\sum_{i=1}^{d} (a_i - b_i)^2\right)^{\frac{1}{2}} = \|\mathbf{a} - \mathbf{b}\|_2$$

$$\left(\sum_{i=1}^{d} |a_i - b_i|^p\right)^{\frac{1}{p}} = \|\mathbf{a} - \mathbf{b}\|_p$$

Similarity measures

What is the distance $d(\mathbf{a},\ \mathbf{b})$ between $\mathbf{a}$ and $\mathbf{b}$?



$\text{distance} = d(\mathbf{a}, \mathbf{b})$
$\mathbf{b} = (b_1, b_2)$
$\mathbf{a} = (a_1, a_2)$
dimension 2
dimension 1

example: $\mathbf{a},\ \mathbf{b} \in \mathbb{R}^2$

- **Manhattan**
  $|a_1 - b_1| + |a_2 - b_2|$
- **Euclidean**
  $\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$
- **Minkowski (p-norm)**
  $\sqrt[p]{|a_1 - b_1|^p + |a_2 - b_2|^p}$

in general: $\mathbf{a},\ \mathbf{b} \in \mathbb{R}^d$

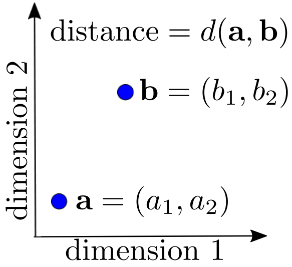
$$\sum_{i=1}^{d} |a_i - b_i| = \|\mathbf{a} - \mathbf{b}\|_1$$

$$\left(\sum_{i=1}^{d} (a_i - b_i)^2\right)^{\frac{1}{2}} = \|\mathbf{a} - \mathbf{b}\|_2$$

$$\left(\sum_{i=1}^{d} |a_i - b_i|^p\right)^{\frac{1}{p}} = \|\mathbf{a} - \mathbf{b}\|_p$$

Similarity measures

What is the distance $d(\mathbf{a},\ \mathbf{b})$ between $\mathbf{a}$ and $\mathbf{b}$?



distance $= d(\mathbf{a},\mathbf{b})$
$\mathbf{b} = (b_1, b_2)$
$\mathbf{a} = (a_1, a_2)$
dimension 2
dimension 1

example: $\mathbf{a},\ \mathbf{b} \in \mathbb{R}^2$

- **Canberra**

$$\frac{|a_1-b_1|}{|a_1|+|b_1|} + \frac{|a_2-b_2|}{|a_2|+|b_2|}$$

- **Cosine similarity**

$$\frac{a_1b_1+a_2b_2}{\sqrt{a_1^2+a_2^2}\sqrt{b_1^2+b_2^2}}$$

- **Correlation distance**

in general: $\mathbf{a},\ \mathbf{b} \in \mathbb{R}^d$

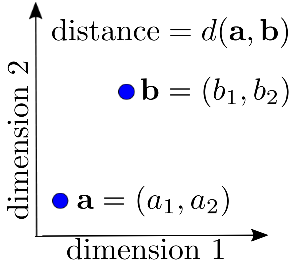
$$\sum_{i=1}^{d} \frac{|a_i-b_i|}{|a_i|+|b_i|}$$

$$\frac{\sum_{i=1}^{d} a_ib_i}{\sqrt{\sum_{i=1}^{d}(a_i)^2}\sqrt{\sum_{i=1}^{d}(b_i)^2}} = \frac{\mathbf{a}.\mathbf{b}}{\|a\|_2\|b\|_2}$$

Pearson (ρ), Spearman, Kendall (τ)

Similarity measures

What is the distance $d(\mathbf{a},\ \mathbf{b})$ between $\mathbf{a}$ and $\mathbf{b}$?


distance $= d(\mathbf{a}, \mathbf{b})$
$\mathbf{b} = (b_1, b_2)$
$\mathbf{a} = (a_1, a_2)$
dimension 2
dimension 1

example: $\mathbf{a},\ \mathbf{b} \in \mathbb{R}^2$

- **Canberra**
  $\frac{|a_1 - b_1|}{|a_1| + |b_1|} + \frac{|a_2 - b_2|}{|a_2| + |b_2|}$
- **Cosine similarity**
  $\frac{a_1 b_1 + a_2 b_2}{\sqrt{a_1^2 + a_2^2}\sqrt{b_1^2 + b_2^2}}$
- **Correlation distance**

in general: $\mathbf{a},\ \mathbf{b} \in \mathbb{R}^d$

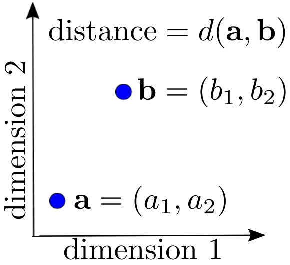
$\sum_{i=1}^{d} \frac{|a_i - b_i|}{|a_i| + |b_i|}$

$$\frac{\sum_{i=1}^{d} a_i b_i}{\sqrt{\sum_{i=1}^{d} (a_i)^2}\sqrt{\sum_{i=1}^{d} (b_i)^2}} = \frac{\mathbf{a}.\mathbf{b}}{\|a\|_2 \|b\|_2}$$

Pearson (ρ), Spearman, Kendall (τ)

Similarity measures

What is the distance $d(\mathbf{a},\ \mathbf{b})$ between $\mathbf{a}$ and $\mathbf{b}$?



distance $= d(\mathbf{a}, \mathbf{b})$

$\mathbf{b} = (b_1, b_2)$

$\mathbf{a} = (a_1, a_2)$

dimension 2

dimension 1

example: $\mathbf{a},\ \mathbf{b} \in \mathbb{R}^2$

- **Canberra**

  $\frac{|a_1 - b_1|}{|a_1| + |b_1|} + \frac{|a_2 - b_2|}{|a_2| + |b_2|}$

- **Cosine similarity**

  $\frac{a_1 b_1 + a_2 b_2}{\sqrt{a_1^2 + a_2^2}\sqrt{b_1^2 + b_2^2}}$

- **Correlation distance**

in general: $\mathbf{a},\ \mathbf{b} \in \mathbb{R}^d$

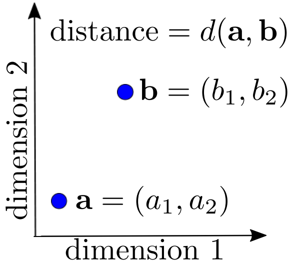
$$\sum_{i=1}^{d} \frac{|a_i - b_i|}{|a_i| + |b_i|}$$

$$\frac{\sum_{i=1}^{d} a_i b_i}{\sqrt{\sum_{i=1}^{d}(a_i)^2}\sqrt{\sum_{i=1}^{d}(b_i)^2}} = \frac{\mathbf{a}.\mathbf{b}}{\|a\|_2\|b\|_2}$$

Pearson (ρ), Spearman, Kendall (τ)

Similarity measures

What is the distance $d(\mathbf{a},\ \mathbf{b})$ between $\mathbf{a}$ and $\mathbf{b}$?



distance $= d(\mathbf{a}, \mathbf{b})$
$\mathbf{b} = (b_1, b_2)$
$\mathbf{a} = (a_1, a_2)$
dimension 2
dimension 1

example: $\mathbf{a},\ \mathbf{b} \in \mathbb{R}^2$

- **Canberra**
  $\frac{|a_1-b_1|}{|a_1|+|b_1|} + \frac{|a_2-b_2|}{|a_2|+|b_2|}$
- **Cosine similarity**
  $\frac{a_1b_1+a_2b_2}{\sqrt{a_1^2+a_2^2}\sqrt{b_1^2+b_2^2}}$
- **Correlation distance**

in general: $\mathbf{a},\ \mathbf{b} \in \mathbb{R}^d$

$$\sum_{i=1}^{d} \frac{|a_i-b_i|}{|a_i|+|b_i|}$$

$$\frac{\sum_{i=1}^{d} a_i b_i}{\sqrt{\sum_{i=1}^{d}(a_i)^2}\sqrt{\sum_{i=1}^{d}(b_i)^2}} = \frac{\mathbf{a}.\mathbf{b}}{\|a\|_2\|b\|_2}$$

Pearson (ρ), Spearman, Kendall (τ)

k-means clustering

```r
fit <- kmeans(x=data, centers=k)
```

1. Select k centroids at random
2. Calculate distance between centroids and each data point
3. Assign each data point to the closest centroid
4. Compute new centroids; the average of all data points in that cluster
5. Repeat steps 2 to 4 until data points remain in the same cluster or some maximum number of iterations reached

k-means clustering

```r
fit <- kmeans(x=data, centers=k)
```

1. Select k centroids at random
2. Calculate distance between centroids and each data point
3. Assign each data point to the closest centroid
4. Compute new centroids; the average of all data points in that cluster
5. Repeat steps 2 to 4 until data points remain in the same cluster or some maximum number of iterations reached

k-means clustering

```
fit <- kmeans(x=data, centers=k)
```

1. Select k centroids at random
2. Calculate distance between centroids and each data point
3. Assign each data point to the closest centroid
4. Compute new centroids; the average of all data points in that cluster
5. Repeat steps 2 to 4 until data points remain in the same cluster or some maximum number of iterations reached

k-means clustering

```r
fit <- kmeans(x=data, centers=k)
```

1. Select k centroids at random
2. Calculate distance between centroids and each data point
3. Assign each data point to the closest centroid
4. Compute new centroids; the average of all data points in that cluster
5. Repeat steps 2 to 4 until data points remain in the same cluster or some maximum number of iterations reached

k-means clustering

```
fit <- kmeans(x=data, centers=k)
```

1. Select k centroids at random
2. Calculate distance between centroids and each data point
3. Assign each data point to the closest centroid
4. Compute new centroids; the average of all data points in that cluster
5. Repeat steps 2 to 4 until data points remain in the same cluster or some maximum number of iterations reached

k-means clustering

```
fit <- kmeans(x=data, centers=k)
```

1. Select k centroids at random
2. Calculate distance between centroids and each data point
3. Assign each data point to the closest centroid
4. Compute new centroids; the average of all data points in that cluster
5. Repeat steps 2 to 4 until data points remain in the same cluster or some maximum number of iterations reached

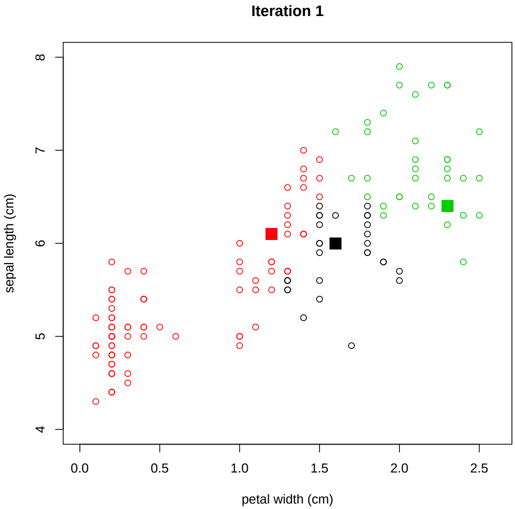
k-means clustering

```
fit <- kmeans(x=data, centers=k)
```

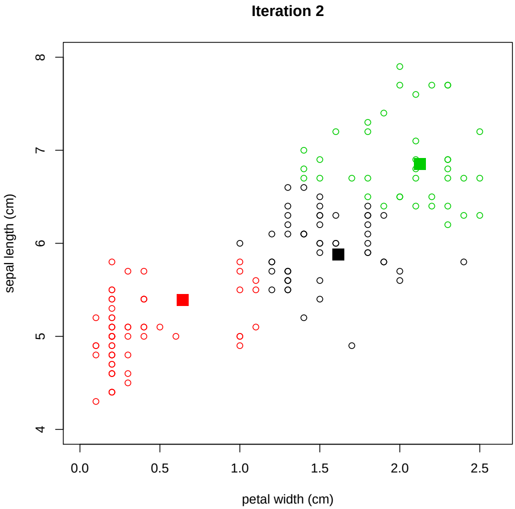
1. Select k centroids at random
2. Calculate distance between centroids and each data point
3. Assign each data point to the closest centroid
4. Compute new centroids; the average of all data points in that cluster
5. Repeat steps 2 to 4 until data points remain in the same cluster or some maximum number of iterations reached

Note: k-means clustering should *only* be used with continuous data

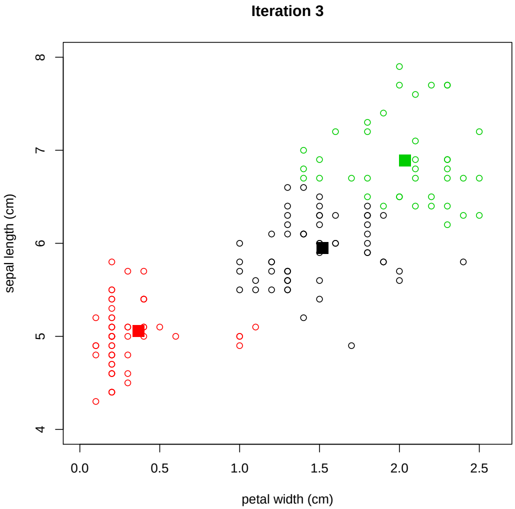
k-means clustering



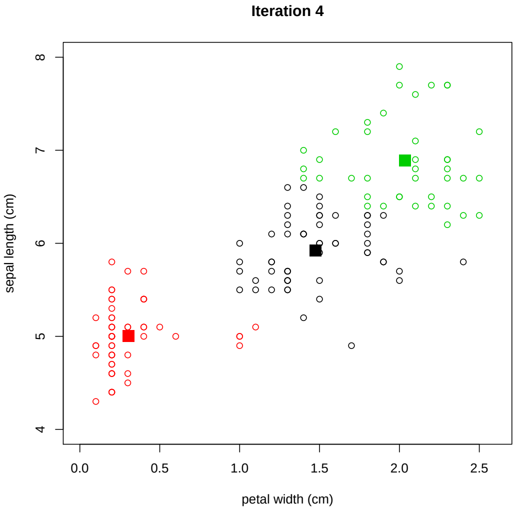
k-means clustering



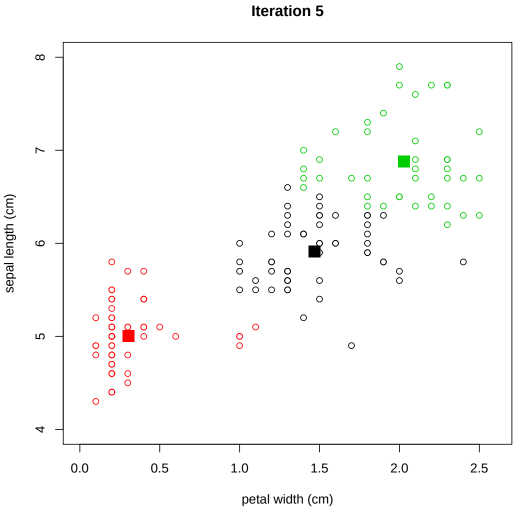
k-means clustering



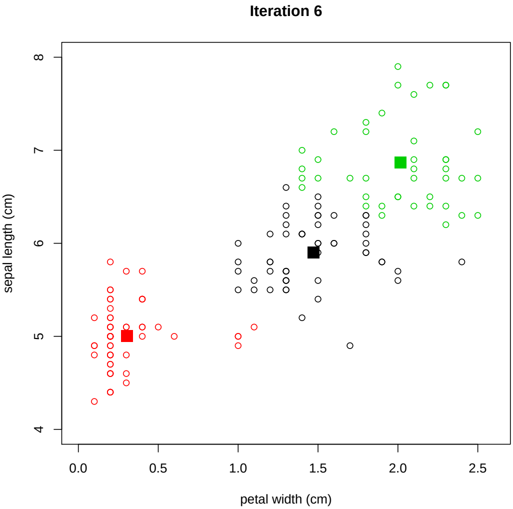
k-means clustering



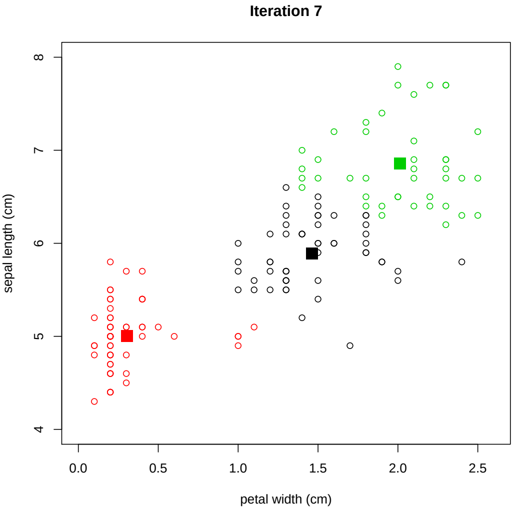
k-means clustering



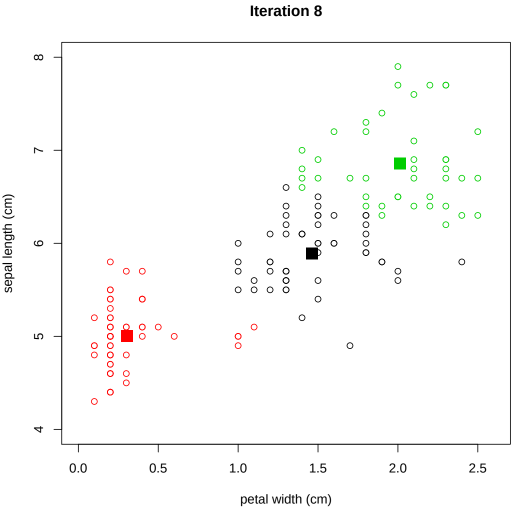
k-means clustering



k-means clustering



k-means clustering



k-means clustering

Pros

- Simple and intuitive
- Computationally inexpensive/fast

Cons

- What is k?
- Only applicable to continuous data where a mean is defined
- No guarantee of a global optimum solution

Agglomerative hierarchical clustering

```
d <- dist(as.matrix(data), method = "euclidean") # Distance method
fit <- hclust(d, method = "complete") # Linkage function
```

1. Assign each data point as its own cluster
2. Compute distance between each cluster
3. Merge the closest pair into a single cluster
4. Repeat 2 to 3 until you're left with one cluster

Agglomerative hierarchical clustering

```r
d <- dist(as.matrix(data), method = "euclidean") # Distance method
fit <- hclust(d, method = "complete") # Linkage function
```

1. Assign each data point as its own cluster
2. Compute distance between each cluster
3. Merge the closest pair into a single cluster
4. Repeat 2 to 3 until you're left with one cluster

Agglomerative hierarchical clustering

```r
d <- dist(as.matrix(data), method = "euclidean") # Distance method
fit <- hclust(d, method = "complete") # Linkage function
```

1. Assign each data point as its own cluster
2. Compute distance between each cluster
3. Merge the closest pair into a single cluster
4. Repeat 2 to 3 until you're left with one cluster

Agglomerative hierarchical clustering

```
d <- dist(as.matrix(data), method = "euclidean") # Distance method
fit <- hclust(d, method = "complete") # Linkage function
```

1. Assign each data point as its own cluster
2. Compute distance between each cluster
3. Merge the closest pair into a single cluster
4. Repeat 2 to 3 until you're left with one cluster

Agglomerative hierarchical clustering

```r
d <- dist(as.matrix(data), method = "euclidean") # Distance method
fit <- hclust(d, method = "complete") # Linkage function
```

1. Assign each data point as its own cluster
2. Compute distance between each cluster
3. Merge the closest pair into a single cluster
4. Repeat 2 to 3 until you're left with one cluster

Agglomerative hierarchical clustering

```
d <- dist(as.matrix(data), method = "euclidean") # Distance method
fit <- hclust(d, method = "complete") # Linkage function
```

1. Assign each data point as its own cluster
2. Compute distance between each cluster
3. Merge the closest pair into a single cluster
4. Repeat 2 to 3 until you're left with one cluster

Note: Step 3 is *key*, the distance method and linkage function dictate the final result

Hierarchical clustering: Link method

How do we calculate the inter-cluster distance? The *linkage function*

Centroid: mean of data points (same as in k-means)

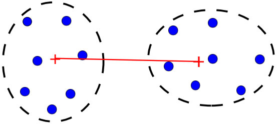
Single: distance between closest pair of points

Complete: distance between furthest pair of points

Average: mean pairwise distance between all points

Hierarchical clustering: Link method

How do we calculate the inter-cluster distance? The *linkage function*



Centroid: mean of data points (same as in k-means)

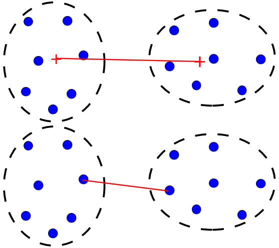
Single: distance between closest pair of points

Complete: distance between furthest pair of points

Average: mean pairwise distance between all points

Hierarchical clustering: Link method

How do we calculate the inter-cluster distance? The *linkage function*



Centroid: mean of data points (same as in k-means)

Single: distance between closest pair of points

Complete: distance between furthest pair of points

Average: mean pairwise distance between all points

Hierarchical clustering: Link method

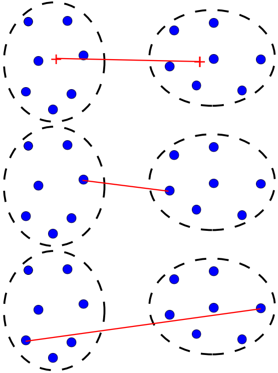
How do we calculate the inter-cluster distance? The *linkage function*

Centroid: mean of data points (same as in k-means)

Single: distance between closest pair of points

Complete: distance between furthest pair of points

Average: mean pairwise distance between all points



Hierarchical clustering: Link method

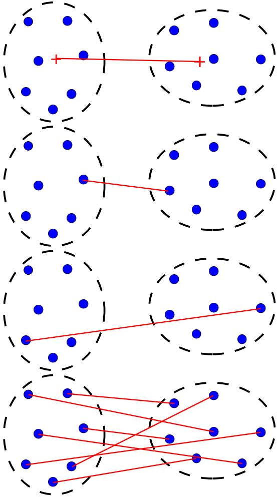
How do we calculate the inter-cluster distance? The *linkage function*

Centroid: mean of data points (same as in k-means)

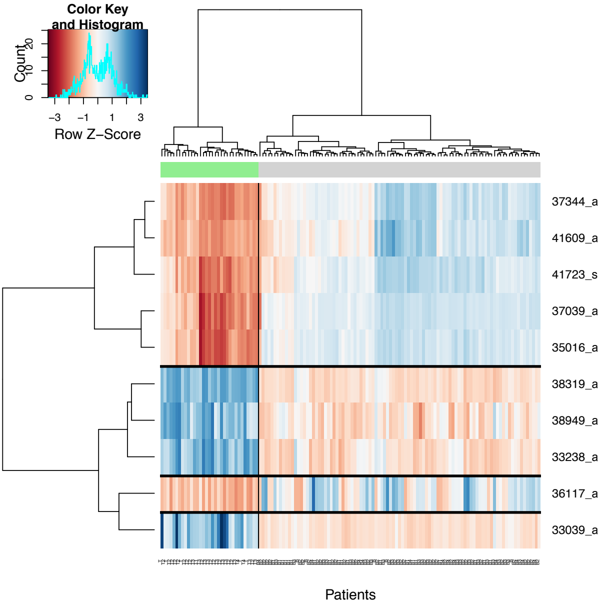
Single: distance between closest pair of points

Complete: distance between furthest pair of points

Average: mean pairwise distance between all points



Hierarchical clustering in gene expression studies



Hierarchical clustering

Pros

- No need to specify k
- Results can be visualised nicely irrespective of number of dimensions

Cons

- Can be computationally expensive
- Interpretation is subjective. Where should we draw the line (to separate clusters)?
- Choice of distance method and linkage function can significantly change the result

Gaussian mixture models

```r
library(mclust)
fit <- Mclust(data, G=k) # G is no. of Gaussians
```

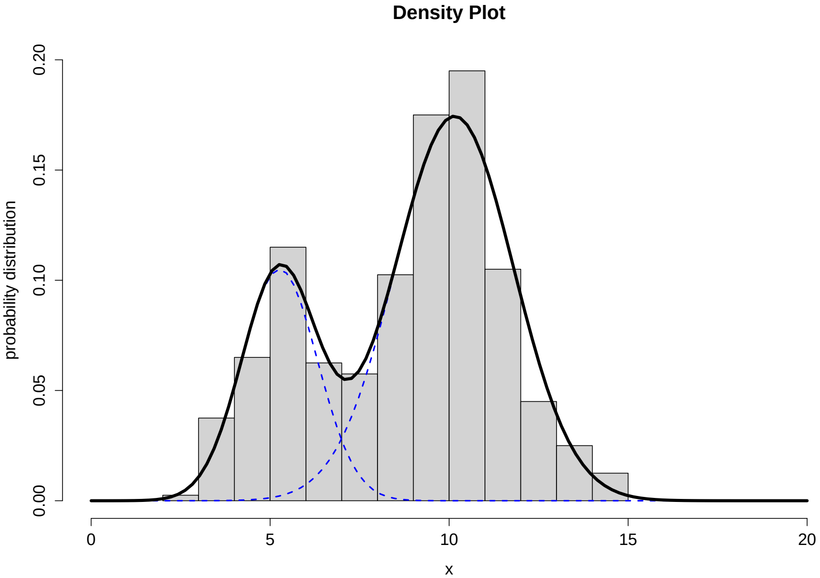
1. Fit k multivariate Gaussian distributions

The Expectation-Maximisation (EM) algorithm is used to estimate the parameters π_i (mixing coefficients), μ_i and σ_i

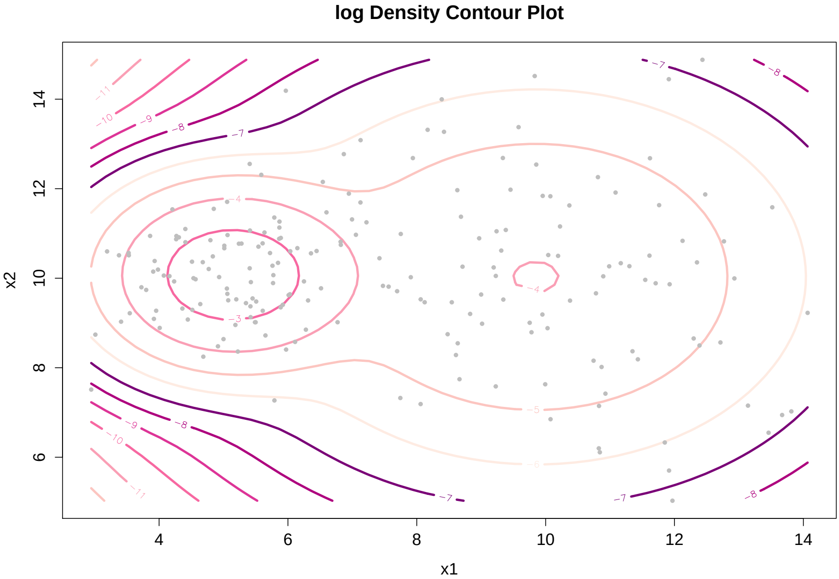
$$p(x) = \sum_{i=1}^{k} \pi_i \mathcal{N}(x|\mu_i, \Sigma_i) \text{ and } \sum_{i=1}^{k} \pi_i = 1$$

Can be seen as a "soft" version of k-means because *every* point is part of *every* cluster but with varying levels of membership

Gaussian mixture models



Gaussian mixture models



Gaussian mixture models

Pros

- Intuitive interpretation
- Computationally inexpensive

Cons

- What is k?
- Strong assumption on the data (normality)
- No guarantee of a global optimum solution

How do we determine the correct number of clusters?

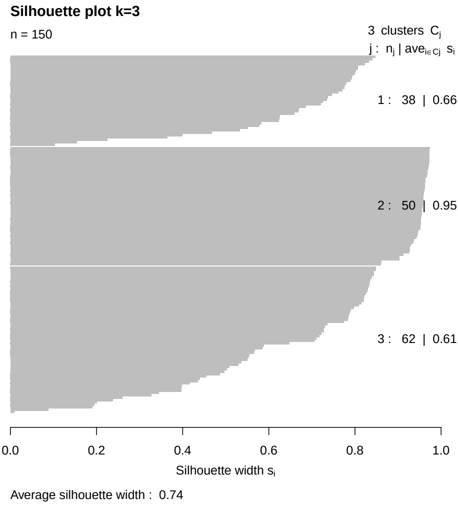
Short answer: you can't

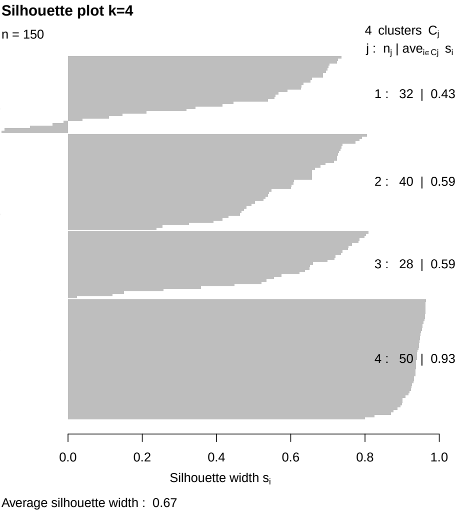
Because data is unlabelled the correct number of k is ambiguous

However we can plot some indices as a function of k to help us evaluate cluster validity:

- Within cluster sum of square distances
- Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) when using distribution-based methods
- Silhouette plot $-1 \geq s(i) \leq 1$ where:
  $s(i) = 1$, *i*th datum is appropriately clustered (good)
  $s(i) = 0$, *i*th datum is borderline between two clusters (meh.)
  $s(i) = -1$, *i*th datum should be in neighbouring cluster (bad)

How do we determine the correct number of clusters?





How do we determine the correct number of clusters?

The NbClust package provides 30 different cluster validity metrics. A majority vote can be taken to deduce the appropriate number of clusters

```r
library(NbClust)
NbClust(data, distance="euclidean", method="kmeans", min.nc=2,
        max.nc=10, index="all")
# distance - similarity measure
# method - clustering algorithm
# min.nc - min number of clusters to consider
# max.nc - max number of clusters to consider
# index - which indices to compute
```

Note: These indices can *only* give us a ballpark range for the correct number of clusters